Extra: Explainable AI (XAI) for deep visual models
Introduction: Post-hoc explanations

- Let Deep models = black box: how to get post-hoc explanations?

- Differs from explainable by design (modularity-inspired models exhibit some forms of interpretability, which can be enforced at different levels in the design of a driving system)
Post-hoc explanations

- Two approaches to explain such models

<table>
<thead>
<tr>
<th>Global explanations</th>
<th>Local explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explain the behavior of a model in general, e.g. across an entire dataset</td>
<td>Given a specific input, justify why the model specifically gives its prediction</td>
</tr>
</tbody>
</table>
Post-hoc global explanations

Prototypes-based methods

- Providing global explanations by selecting and aggregating multiple local explanations, ie find prototypes (specific data instances representing well the data) & criticisms (instances not well represented by the set of prototypes) and see the model predictions on these examples
Local explanations

Given this situation $x$, the output decision taken by the network is $y$.
Why?

- Input **saliency** visualization = **input attribution**
- **Counterfactual** interventions = inferring the prediction of a model for imaginary inputs that have not been observed
XAI by Saliency
Local explanations: input attribution

- **Attribution methods**: identify the most relevant parts considered by a neural network for a given class
Local explanations: input attribution

\[ X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,w} \\ \vdots & \ddots & \vdots \\ x_{h,1} & \cdots & x_{h,w} \end{bmatrix} \]

\[ F = g_l \circ g_{l-1} \circ \cdots \circ g_1 \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_C \end{bmatrix} \]

\[ F(X) = y \]

> Attribution method: identify which subset of variables (pixels) \( X \) that has the greatest impact on a specific output \( y_c \)
Local explanations: input attribution

- It depends on how the "importance" of variables is defined

\[ F(X) = y \]

- Gradient-based importance:
  - Vanilla gradient: \( \| \frac{\partial F}{\partial X_i} \| \)
  - Modified gradient: Guided Backpropagation, Deconvolution

- Other popular methods:
  - Grad-CAM: Spatial averaged gradient * value \( \left( \frac{1}{N} \sum n \frac{\partial F}{\partial X_n} \right) X_i \)
# Local explanations: Input saliency visualization

<table>
<thead>
<tr>
<th></th>
<th>orig img + gt bb</th>
<th>gradient</th>
<th>guided</th>
<th>contrast excitation</th>
<th>grad-CAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>chocolate sauce</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td>Pekinese</td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td>cliff</td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td>street sign</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Local explanations: Input saliency visualization

Local explanations: Input saliency visualization

Bojarski et al. 2017

Liu et al. 2019
Local explanations: Local approximation

Forward approach (No gradient): explain the behavior of the black-box model in the vicinity of the instance with a simpler model.

Figure 4: Explaining an image classification prediction made by Google’s Inception neural network. The top 3 classes predicted are “Electric Guitar” \((p = 0.32)\), “Acoustic guitar” \((p = 0.24)\) and “Labrador” \((p = 0.21)\).
Local explanations: Input saliency visualization

Useful to analyze and detect **spurious correlations**

(a) Husky classified as wolf  (b) Explanation

[ribeiro2016lime]

[Scenario A: Full Information]
- Policy attends to brake indicator
- Brake

[Scenario B: Incomplete Information]
- Policy attends to pedestrian
- Brake

[Causal Confusion in Imitation Learning, S. Levine NeurIPS19]
Interpretability/Explainability vs. Performance

Trade-off between model interpretability and performance, and a representation of the area of improvement where the potential of XAI techniques and tools resides.
XAI by counterfactual analysis
Counterfactual visual explanations (Goyal et al., ICML 2019)

- Using an input image $X$ classified as $c$, and $X'$ classified as $c'$
- Find which activations of $X$ should be replaced by activations of $X'$, to be classified as $c'$
Counterfactual explanations (for classifiers)

DiVE (Rodriguez et al., ICCV 2021)

- Variational autoencoder, to encode images into disentangled latent vectors $z$
- Optimize a set of perturbations on the latent vector, to switch the decision of the classifier
Generation of counterfactual samples

Original Image $I$

Semantic mask $S_I$

Encoder $E$

Style codes $z^I$

Counterfactual codes $z$

Generator $G$

Counterfactual Explanation $I'$

Discriminator $D$

Victim Model $M$

$L_{adv}$

$L_{decision}$
Qualitative results (CelebAMask-HQ)

When $y' = 0.5$ (i.e. we aim at border of the decision):

Reconstructed Image
- Male score: 0.99
- Young score: 0.97

Why is X a male and not a female?
- Counterfactual explanation
  - Male score: 0.23

Why is X young and not old?
- Counterfactual explanation
  - Young score: 0.49
Qualitative results (CelebAMask-HQ)

When \( y' = 0.0 \) (i.e. we aim at entirely switching the decision):

Why is X a male and not a female?

Counterfactual explanation
Male score: 0.009

Reconstructed Image
Male score: 0.99
Young score: 0.97

Counterfactual explanation
Young score: 0.002

Why is X young and not old?
Qualitative results (BDD)

Real Image $I$

$M(I) = \text{Go forward}$

$M(I') = \text{Stop}$

$M(I') = \text{Go forward}$

Countfactual Explanation $I'$

$M(I) = \text{Go forward}$

$M(I') = \text{Stop}$

$M(I') = \text{Go forward}$

$M(I) = \text{Stop}$
Qualitative comparison (BDD)

Real image
Decision: do not go forward

Counterfactual explanation
Decision: go forward